DOCUMENT RESUME

ED 392 846 TM 024 694

AUTHOR Kahl, Stuart R.
TITLE Scoring Issues in Selected Statewide Assessment Programs Using Non-Multiple-Choice Formats.
PUB DATE 18 Apr 95
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Assessment; *Equated Scores; *Performance Based Assessment; *Portfolios (Background Materials); Psychometrics; *Scoring; *State Programs; Test Construction; Test Format; *Testing Problems; Testing Programs; Test Reliability
IDENTIFIERS *Alternative Assessment; Scoring Rubrics

ABSTRACT

Although few question the positive impacts alternative forms of assessment can have on instruction, concerns about the psychometric quality of data obtained from such assessments are taking their toll. Scoring issues are at the heart of many of these concerns. This paper addresses the causes of these concerns: misinformation about psychometric quality of different modes of performance assessment and resistance to applying existing, sound measurement practices in some performance-based assessment programs. Specific topics addressed include: (1) reliability of open-response tests; (2) double-scoring versus using more items; (3) unique problems of portfolios; (4) characteristics of scoring rubrics; (5) general scoring guides; (6) questions that do not allow "entry" for all students; (7) quantitative scoring rubrics; (8) equating issues; (9) giving scoring information to students; and (10) issues in writing scoring rubrics. The four exhibits include a general scoring guide and sample test items with scoring rubrics. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 392 846

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
STUART R. KAHL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Scoring Issues in Selected Statewide Assessment Programs Using Non-Multiple-Choice Formats

Stuart R. Kahl

Advanced Systems in
Measurement and Evaluation, Inc.
Dover, New Hampshire

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1995.

TM024694

BEST COPY AVAILABLE

Introduction

The March 27, 1995 issue of *Education Week* included an article entitled "The New Breed of Assessments Getting Scrutiny" (Olson, 1995). That article suggested that there are "signs of retreat" from large-scale performance assessments. It would be unfortunate indeed if such a retreat were sustained. There are few who question the positive impacts alternative forms of assessment can have on instruction. But concerns about the psychometric quality of data obtained from such assessments are taking their toll. There are two general reasons for these concerns: misinformation about the psychometric quality of different modes of performance assessment and a reluctance to apply existing, sound measurement practice in some performance-based assessment programs. At the heart of many concerns regarding alternative assessments are scoring issues.

Misinformation about Data Quality

The *Education Week* article reports that researchers question the reliability of Kentucky's assessment system and "have cautioned against using results from the performance events or writing portfolios to make high-stakes decisions because of problems with reliability." It is true that these two types of assessment, by themselves, currently are not sufficiently reliable for high stakes decisions. However, what the article does not point out is that the cognitive index computed for Kentucky schools is based primarily on results from on-demand open-response tests, and therefore the index is reliable. Adding the less reliable components to the composite with limited weight makes the index slightly less reliable. Whether a slight reduction in reliability coefficients (still maintaining them at a certain level) by including less reliable, but consequentially valid, components is inappropriate has not been established.

For some reason, the literature on alternative assessments over the past five years or so, has focused on extended performance events and portfolios, leaving on-demand open-response testing somewhat of a "forgotten format." The failure to distinguish among the different modes of performance assessment (open-response, extended performance events, and portfolios) results in potential users of non-multiple-choice tests being misinformed about the quality of data that can be obtained from them. As a consequence of misinformation, the retreat discussed by *Education Week* is a retreat back to the multiple-choice format. The assessment programs in states that have moved more cautiously than others by emphasizing on-demand open-response questions are struggling for their existence because of the bad press given to performance assessments generally, despite the high psychometric quality of this form of assessment.

Resistance to Sound Measurement Practice

The *Education Week* article quotes one distinguished educator as saying, as many have, that advocates of performance assessment have "moved out in front of the psychometric technology." Perhaps a more accurate description would be that many advocates of performance assessment, in their zeal to rebel against the use of traditional measures, have resisted the use of sound measurement

practices in conjunction with their alternative forms of assessments. The portfolio components of assessment programs in Vermont and Kentucky illustrate this. In the early developmental stages of Vermont's program, curriculum specialists comprising the advisory committees went so far as to resist the appearance of numerals associated with descriptions of levels of performance on the scoring guides. They admittedly were trying to forestall any efforts to aggregate data despite the fact that the needs for statewide accountability data were the justification for the funding of the portfolio assessment. Years later, evaluators of the program from the Rand Corporation suggested the need for greater restrictions in the types of tasks included in the portfolios (Koretz et al, 1994). Such restrictions, including the use of the some common tasks which could serve to anchor the scoring, were strongly resisted by the content advisory committees in Vermont.

In Kentucky, similar restrictions have yet to be adopted. Also, in avoiding any tendency to be overly analytic, Kentucky advisory committees have steadfastly endorsed holistic scoring of the portfolios, which yields a single rating of each portfolio on a 4-point scale. As far as the computer is concerned, there is no difference between a portfolio scored from 1 to 4 and a half-page response to a single open-response question scored from 1 to 4. In terms of reliability, the computer is right.

Such findings should not be surprising. The challenge to advocates of performance-based assessment is not to survive while psychometricians make some new, revolutionary discovery that will solve the problems of performance events and portfolios. That is not going to happen. The challenge is to use what is already known about measurement to make such modes of assessment reliable.

As stated earlier, scoring issues are at the heart of data quality. Since experience is the best teacher, statewide assessment programs, which have led the way in the implementation of alternative forms of assessment, should be a valuable source of lessons. In the remainder of this paper, some familiar maxims are interpreted in relation to scoring issues and to lessons learned from selected statewide assessment programs. The second half of the paper gets down to the "nuts and bolts" level in its discussion of scoring rubrics.

Throughout this paper, reference will be made to open-response questions. The reader should assume that the discussion in this paper refers to questions that require five to ten minutes to answer and approximately a half page of response space. These produce responses that are substantial enough to be scored on a 0-to-4 scale, yet short enough that enough questions can be asked in a reasonable testing time to provide adequate generalizability of results. The scoring approach discussed is not the only scoring approach, but it appears to be an effective and efficient approach for large scale assessment programs in which scoring costs must be carefully controlled.

WHEN I (A TEST QUESTION) AM GOOD, I'M GOOD; WHEN I'M BAD, I'M BETTER

The Reliability of Open-Response Tests

Multiple-choice tests have been called objective tests. Therefore, people assume open-response tests must be subjective. It is true that scorers of responses to open-response questions must make judgments. However, every action taken in conjunction with the scoring process in a large-scale assessment program is designed to make those judgments as objective as possible. Those actions include the preparation of a unique scoring rubric for every question, the use of training and qualifying packets consisting of actual student responses, the training itself, and the continual monitoring of scorers' work. If total objectivity and subjectivity are the extremes of a continuum, the scoring of open-response questions is nowhere near the subjectivity extreme. Note, however, that while the scoring of multiple-choice questions is clearly more objective than the scoring of open-response questions, the inferences about student capabilities based on multiple-choice measures are far less direct.

The notion that an open-response test is inherently less reliable than a multiple-choice test is just plain false. We can make an open-response test as reliable as we want simply by increasing the number of questions; the number of such questions (scored on a scale from 0 to 4) needed to match the reliability of a typical multiple-choice test is approximately one-fourth the number of multiple-choice items. Thus, an open-response question is roughly equivalent to four multiple-choice questions in terms of its contribution to test reliability. The ratio of 1 to 4, by the way, is not just coincidence; it relates to the number of points on a performance continuum at which the items discriminate — four for open-response questions, one for multiple-choice questions.

Is the scoring of responses to open-response questions perfect? No. However, educators are mistaken when they equate scoring inaccuracies with measurement error or scoring consistency with test reliability. They are not the same. Suppose a student's response was incorrectly scored. For instance, suppose a response was awarded 3 points, when it should have been given 4 points. Three points on a response that could have earned 0, 1, 2, 3, or 4 points tell us much more about the student than 1 point on a multiple-choice item with the possible scores of only 0 and 1. With multiple measures and scoring errors balancing out, it takes many fewer open-response items to achieve the same reliability as a much longer multiple-choice test (i.e., longer in terms of number of items).

Returning to the maxim, many people consider scoring error/inconsistency to be bad. However, since we can obtain the same level of reliability with open-response questions that we can with multiple-choice questions, and since the open-response test almost certainly has greater consequential validity, then bad must be better.

IT IS (SOMETIMES) BETTER TO DO SOMETHING RIGHT THE FIRST TIME THAN TO HAVE TO DO IT AGAIN.

Double Scoring vs. More Items

Concerns about the accuracy of scoring and its impact on reliability often lead to the suggestion that pieces of student work should be scored more times. There are some situations in which multiple scorings may be appropriate. However, twice the number of questions scored once is more reliable than half the number of questions scored twice. When there are, by necessity, a limited number of measures within a test, then double scoring is frequently advisable. For example, writing samples solicited by a single writing prompt and producing individual student writing scores should be scored at least twice. If there are high stakes attached to whether students exceed or fail to exceed a particular score, then many more scorings would be appropriate. (Some efficiency is achieved if only the writing samples from students scoring near the cut point are scored more than twice.) However, in the case of open-response questions in programs with higher stakes for schools and lower stakes for individual students, it is often better to broaden the test by asking each student to respond to more questions. The table below shows the effects on reliability of more scorings versus more items.

Coefficient Alphas Based on Double and Single Scoring of Responses to Extended Open-Response Questions

	6 Questions		12 Questions	
	Single	Double	Single	Double
Grade 5 (n = 155)				
Reading	.70	.72	.83	.86
Mathematics	.67	.67	.74	.75
Grade 7 (n = 168)				
Reading	.82	.85	.87	.89
Mathematics	.69	.70	.81	.82
Grade 10 (n = 154)				
Reading	.82	.85	.90	.91
Mathematics	.71	.74	.81	.83

NOTE: data from local district testing program; scoring by teams of 13 to 14 "off-grade" teachers

Clearly, in terms of reliability, more questions are better than more scorings. Unfortunately, there are other factors that have to be considered besides reliability. Even if an assessment's primary focus is school program evaluation, if it produces student level results, there are reasons to consider double scoring. Despite the high reliability that can be obtained with single scoring, there will be scoring errors made on the responses of individual students. The students, their parents, and their teachers do not accept such errors readily. This is especially true if the responses, the questions and the scoring guides are released to the schools. One possible compromise solution to this problem would be to double score responses to common questions that all students answer (these are generally used to produce student results) and to single score responses to matrix-sampled questions (questions that vary across forms of the test and broaden the coverage of a content domain to produce more generalizable school level results). However, when responses are scored multiple times, scoring costs increase considerably, unless of course, the number of questions per student is reduced.

YOU CAN'T HAVE YOUR CAKE AND EAT IT, TOO.

Unique Problems of Portfolios

Educators have been very forgiving when it comes to the reliability of a single writing prompt. If the responses are scored holistically, the prompt constitutes a one-item test. In response to those who are critical of the reliability of such measures, portfolios have been proposed. The thinking here is simply that portfolios contain multiple entries; therefore there are multiple measures. This is NOT the case if a portfolio is given a single holistic rating, double scored or not. The issue is not simply one of scoring accuracy; it is also a matter of reducing data to too limited a scale too early. In the case of portfolios in some programs, the scorer makes the final, direct determination of a student's performance level. There are good reasons for the current trend of reporting assessment results in terms of a limited number of performance levels. However, even with no misclassifications, the low students in one level are more like high students in the next lower level than they are like high scoring students in their own level. Thus, very fine distinctions are required for students near the cut points. Holistic judgements must produce more misclassifications than a more substantial, reliable measure for which cut points are determined.

Accepting the need for "real" multiple measures from portfolios, we then must answer the question of how to obtain multiple measures. Is scoring the individual entries separately the answer? Probably in writing only. Writing is probably a generalizable enough skill that common criteria could be applied across entries. In other disciplines this is not likely the case. An alternative to scoring entries was used in Vermont's mathematics portfolio assessment. In that program, seven different attributes of students' work were evaluated; these attributes pertained to aspects of problem solving and communication. However, criteria for these attributes (multiple measures) did not apply equally as well across portfolio entries. The hope in Vermont is that over time teachers would be able to work with students to include entries to which the criteria can be applied equally as well.

This may not be possible — not because of any limitations of the teachers, but rather because it simply may not be possible at all, given the nature of the criteria.

For mathematics there may be an alternative approach to obtaining multiple measures which does not require that criteria apply equally as well across entries. Suppose the domain of mathematics is subdivided into a limited number of subdomains such as the familiar ones of (1) numbers, numeration, and procedures; (2) geometry and measurement; (3) probability and statistics; (4) variable and relationships. Suppose further that performance levels are defined for each of these subdomains. For example, in geometry and measurement, part of the definition of the novice level might indicate that students could read instrument scales, and identify appropriate units. . . . Distinguished students may be expected to develop an approach to evaluating performances of some kind by quantifying attributes of the performances so that their quality can be measured. They might also be expected to demonstrate understanding of measurement error and precision. With performance levels so defined for each of the subdomains, the teachers may then be charged with assisting each student in including as portfolio entries pieces of work that demonstrate the student's highest level of proficiency in each subdomain. Each entry would not have to address each subdomain, but an entry could address more than one. This is not so different an expectation of teachers involved in portfolio assessment already, but it would make their task of assisting students clearer and more systematic. Banks of sample tasks could include information on the mapping of tasks to levels within subdomains and thus further assist teachers in guiding students toward appropriate activities and entries. Teachers and students could also include in the portfolios introductory material (they already do) that directs scorers to different entries for evidence related to the different subdomains. The scorers of course, would assign a rating for each subdomain. The process would yield multiple measures, based on common criteria that could be applied to portfolios in which entries vary from student to student. Scorer agreement would be enhanced by the more specific description of performance levels within subdomains. The subdomain scores could be aggregated later and cut scores or rules for determining total math performance levels established. The process would also generate more diagnostic information at the school and student levels than current approaches produce. The lack of such information has been a concern many educators have had about the results from alternative forms of assessment.

While an approach such as the one described above could improve scoring consistency, there would still be a need for an audit system in large programs in which teachers score their own students' portfolios. While the portfolio audit system in Kentucky continues to be refined, the experience of the 1992-93 audit process provides some interesting lessons. That year, 105 schools were identified whose writing portfolio scores were most discrepant with those schools' own results on other measures, including an on-demand writing test and the previous year's reported portfolio scores. All of the portfolios from those schools were rescored out of state, along with a random sample of portfolios collected from schools statewide. While the scores from the audit showed that there was considerable misscoring of portfolios statewide, the misscoring in the 105 schools purposively selected was far greater. After the audit results were returned to the schools, the school officials were given the option of submitting all portfolios for which the teachers' scores and the audit scores differed to a scoring team of Kentucky teachers for a third scoring. The results of the audit scoring are shown in the table below.

Numbers of Portfolios Assigned to Each Scoring Level As Determined by Kentucky Teachers in Audit Schools, the New Hampshire Audit Team, and the Review Team of Kentucky Teachers

	Group Doing the Scoring		
Score Assigned	Kentucky Teachers in Audit Schools	New Hampshire Audit Team	Review Team of Kentucky Teachers
Incomplete	6	2	12
Novice	38	464	470
Apprentice	328	203	178
Proficient	268	33	38
Distinguished	62	0	4
Average Accountability Index *	69.2	16.3	16.4

* on a scale from 0 to 140.

The data reported above, of course, shows that a high degree of consistency can be obtained from central scoring at different sites with different scorers. What it does not show is what happened as an aftermath to the 1992-93 audit experience. Representatives of the audited schools were invited to regional meetings to discuss the portfolio scoring and audit processes and results. A general conclusion of the participants was that the primary reason for the discrepancy between original scores and audit scores was that the teachers' scoring focused on a small portion of the scoring criteria — primarily the sections related to writing mechanics (spelling, grammar, punctuation), which tended to be well done in the portfolios, as might be expected. The well-trained scorers involved in the audit scoring were correct in assigning the lower scores. Furthermore, the following year, the staffs of the audited schools demonstrated considerably greater accuracy in scoring than school staffs statewide. Thus, as a result of the audit process, the teachers in the audited schools understood the scoring criteria far better than other teachers. This also resulted in improved performance the next year.

Returning to the issue of multiple measures, it is obvious from a measurement perspective, that more measures are better than fewer measures. This applies whether one is dealing with on-demand open-response tests, extended performance events, or portfolios. In Kentucky, every student in grade 4, 8 or 11 participates in performance events. These "hands-on" activities take about 45 minutes to complete. They involve the students in small group work first; then they require students to work individually to produce scorable products. In the past, at grade 4, the products have

received a single score from 0 to 4. At the other grades, various components of the scorable products received separate scores. Needless to say, the situation at the upper grades is more desirable. This becomes particularly evident when trying to equate instruments across years. Collapsing performance event scores to performance levels is only reasonable when one starts with scores representing a much broader continuum.

One final word about portfolios and performance events as described above is in order. Because the products of these modes of assessment are often the result of some cooperative work, the question is often asked, "Whose work is this anyway?" Requiring some entries that are the result of on-demand testing with no cooperative work has been suggested as a way of dealing with the problem. However, unless all entries are produced this way, there will always be "contamination." Perhaps it is best to recognize that in real life, we work cooperatively some times, and at other times we do not. Therefore, portfolios or performance events representing cooperative work may be a legitimate component of an assessment program, but they should certainly not be the only component. We need to know what every child can do on his or her own as well.

FOOLISH CONSISTENCIES ARE THE HOBGOBLINS OF TEST DEVELOPERS

Characteristics of Scoring Rubrics

Open-response questions, performance-events, even portfolios — none of these are new developments. Yet decades of heavy reliance on multiple-choice instruments have left many educators feeling insecure about their abilities to use these alternative forms of assessment. Teachers trying to change the way they test, state department staffs trying to create new assessments, and contractors trying to help them are all looking for simple, formulaic rules for developing questions and scoring rubrics. For getting started, perhaps such rules are necessary, but in the long run they might be detrimental.

General Scoring Guides

A fairly common tool test developers use to guide their development of scoring rubrics for specific test questions is a general scoring guide. Some have suggested that definitions of performance levels, with respect to which test results may ultimately be reported, should serve as the general guide for developing specific rubrics. This approach has some problems. First, just as multiple-choice questions discriminate at different levels, so too do score points for open-response questions. For example, the 4-point response for one question may cut off the top 5 percent of students along the ability continuum, while the 4-point response to another question may cut off the top 20 percent. It would not be possible to force open-response questions to discriminate identically. Furthermore, since it is better to assign students to performance levels based on multiple measures, we should not confuse people by referring to a "distinguished" response or a "proficient" response to a single question. We should refer to a "4-point response" or a "3-point response" and make it clear that the other terms refer to performance levels based on total test results.

Another general scoring guide, much like ones used in several states, is shown in Exhibit A, attached to this paper. Developing item-specific rubrics to be consistent with such a guide seems logical, but the general guide can cause problems as well. By design, it has a "holistic feel" to it and is not item-specific. People using this general guide tend to produce item-specific rubrics that also have a "holistic feel" to them and do not seem very item-specific. For example, consider a scoring rubric for a question asking students to describe the character of Lewis using evidence from an excerpt from "Deliverance" to support the description. A fairly ineffective rubric for this question might look like the following:

4 points: provides in-depth analysis of Lewis
3 points: provides reasonable description of Lewis with support
2 points: identifies a likely attribute of Lewis
1 point: provides a literal statement about Lewis taken directly from the story

The distinctions between some adjacent score points in the guide above are not clear. A better guide might make distinctions between important versus superficial attributes of Lewis, place higher value on multiple pieces of evidence from the excerpt, and address the "fit" of evidence provided to the attribute(s) named. Brief examples (words and phrases for attributes or support) could be embedded in the score point descriptions without making the rubric too long.

Sometimes a "holistic" approach to a scoring rubric makes scoring difficult because it is inconsistent with the structure of the question. For example, some multipart questions may be more efficiently and accurately scored by dividing the points available across parts and having each part rated on a 1- or 2-point scale. The mathematics item and scoring rubric shown in Exhibit B illustrate this approach, although the same approach could be used for questions in other disciplines.

Questions That Do Not Allow "Entry" for All Students

As greater use is made of higher-order, constructed-response questions, some new problems emerge. This appears to be particularly true in mathematics. Some complex or nonroutine problem solving situations may not "provide entry" into the problem for all students. A good open-response question discriminates at several levels spanning the full range of the ability continuum. However, if some students cannot identify an appropriate solution strategy for a problem, they may not be able to show on that problem any of the proficiencies they actually do have. High nonresponse rates result. On questions in other disciplines, this seems to be less of a problem. Two approaches to dealing with the problem in mathematics involve the structure of the questions and scoring. So-called "scaffolded" questions can be helpful. These are multipart questions, the parts of which are increasingly more sophisticated, thus forming almost a Guttman scale. The first part, of course, should be doable by relatively weak students. For example, data on taxi cab fares for different trip distances could be provided to students. The first part could ask the students to plot the data on a graph. The second part could ask the student to estimate the fare for a distance not represented in the data provided. (Students could answer this part by interpolating or extrapolating from the graph or by using algebra.) Finally, students could be asked to explain the significance of the slope and y-intercept in the equation corresponding to the graph. Each part of such a question could be worth one or two points, with the item score simply being the sum of the part scores.

A second approach to the "entry problem" in mathematics might be to give credit (e.g., one point) for any correct mathematics even if it is totally inappropriate in terms of being a step toward the ultimate solution of the problem. In other words, if a student does not understand a problem enough to choose an appropriate strategy, but makes an attempt and correctly computes, say, a percentage, then the student would be awarded a point. After all, students at lower levels of proficiency generally have some skills, even though they may have difficulty knowing when to apply them.

"Quantitative" Scoring Rubrics

Quantitative scoring rubrics are rubrics that rely heavily on counts of response components. In Kentucky a few years ago, a statement was made to content advisory committees involved in test development that quantitative scoring rubrics make the describing of performance levels difficult. It is better to describe qualitative differences between performances of students at different levels, that is, qualitative differences between responses typical of the different score points for items. Thus, in the early years of the Kentucky program, it was advisable to avoid quantitative scoring guides to facilitate the understanding of the performance levels. As a result, a message was passed throughout the state that quantitative scoring rubrics are bad. This is not true. In scoring responses to a question we want to use whatever is the best means to separate stronger responses from weaker responses. Sometimes quantitative approaches work well. The math questions discussed previously (one in Exhibit B and the example of a scaffolded item) have what might be considered quantitative scoring rubrics calling for scorers to count the number of correct responses to parts. Sometimes simply counting examples given in response to questions in other subjects is effective. For example, a science question might describe a simple electrical circuit (including a battery, a bulb, a switch, and wires), then ask the students to give as many reasons as they can why the bulb might not light up. The scoring rubric for the item could list the reasons and require the scorer to simply award a point for each one named. (This is not a low-level question for intermediate level students.) Furthermore, such a quantitative rubric can be considered consistent with the general scoring guide shown in Exhibit A. The student who can give more reasons in response to a particular question may well have more in-depth understanding or show greater insight.

Quantitative scoring guides that might be worth avoiding would be ones that are too complicated for scorers to use efficiently. For example, if the description for a score point reads "a strong conclusion for Part A with two good examples and a weak explanation for Part B with one or two examples OR a weak conclusion for Part A with two or more good examples and a strong response to Part B with two good examples OR . . . ." In other words, if there are too many combinations to describe simply, then it is probably better to use another approach -- e.g., two points per part. This does not mean to say, "NEVER use the 'combinations' approach." Sometimes, there could be fewer combinations to include in a score point descriptor and the approach would work well. It depends on the structure of the item and on what the item is intended to measure.

Equating Issues

There may be other problems associated with the use of a general scoring guide. For example, consistently scoring responses to the same questions administered in successive years of

a program appears to be a difficult task. This means that in equating tests over time, both changes in performance and changes in the way responses are scored must be taken into account. The second factor is not an issue in multiple-choice testing; but in open-response testing it necessitates the "seeding" of old responses in with new responses and special analysis of the differences between their original scores and scores from rescoring. Some testing specialists suspect that the use of a general scoring guide may contribute to the problem. Especially in situations in which professional scorers move from one project to another, one subject to another, even one grade to another, there is the concern that they so internalize a general scoring guide that they apply common holistic criteria across projects, subjects, or grades. Over time their standards drift, but in an uncontrolled manner, not because of the uniqueness of the scoring task at hand. This problem contributed to NAEP's decision to use primary trait instead of holistic scoring of writing samples. It is important that scorers change their standards appropriately to match the specific standards associated with each specific scoring rubric for each particular prompt or test question.

Giving Scoring Information to Students

Another interesting problem regarding a general scoring guide has arisen in Kentucky. It seems reasonable to let students know how their work is to be scored. Yet sharing item-specific scoring rubrics would give away answers. Consequently, Kentucky test forms include a general scoring guide to assist students. Unfortunately, as we can see from Exhibit A, the description of a 4-point response lists several characteristics of responses that may not apply to all test questions. Yet, when teachers tell their students that these characteristics are what the scorers are looking for in every response, they deliver the message that students should write everything they know about the topic area of each question, whether it is asked for or not. The result is obvious. Students are including a great deal of irrelevant discussion in their answers. Scoring time is increased dramatically, and scorers have to "dig" to find out if the students understand the concepts being tested. The added information may do more harm than good for the students.

YOU GET WHAT YOU ASK FOR
OR
YOU GET WHAT THEY THINK YOU ARE ASKING FOR
OR
WHAT THEY SEE IS WHAT YOU GET

The last problem discussed above is typical of communication problems that occur in conjunction with high stakes testing. There are other examples of such communication problems related to scoring practices. For example, earlier in the section above, the overgeneralizing of the statement about quantitative scoring rubrics was discussed. This was a communication problem. The Kentucky program provides still more examples. The first year of Kentucky's program, scorers encoded one of the following scores for every response to an open-response question: 0,1,2,3, or 4. "Zero" was the code for a blank (nonresponse), and "1" represented responses that were totally wrong or showed only minimal understanding. These scoring codes communicated to teachers and students that "you can get a point just for writing anything in an answer space." The impact of this

message the following year was obvious. Fortunately, because cut scores separating the bottom two performance levels in the different subject areas were considerably higher than points on the total test score continua corresponding to all or mostly 1's, this problem only affected the performance level designations of a handful of students. Nevertheless, the following year the scorers had an additional empty bubble to mark when there was nothing at all written by a student in an answer space. This left the "zero" for totally incorrect or irrelevant responses.

A similar problem resulted from the public release of questions, scoring rubrics, and sample student responses. In selecting sample responses associated with different scores on an item, the samples that were chosen to represent low scoring responses were considerably shorter than those representing the strong responses. This relationship between quality and length was typically the case. Unfortunately, the message received was, "to get higher scores, just write more." That message, along with the relaxing of restrictions on testing time and response space, has resulted in the production by many students of a great deal of text with little substance. Scoring time has increased dramatically. Scorers believe that students are getting lower scores than they otherwise might have because they are writing too much and the correct portions of their answers are lost in a sea of irrelevancies. This matter is currently being investigated; information on length of responses is being gathered during scoring.

BREVITY IS THE SOUL OF WIT
OR
BREVITY IS NEXT TO CLARITY

There is sometimes a tendency for developers of item-specific scoring rubrics to produce lengthy descriptions of responses associated with the different score points. This practice should probably be avoided whenever possible. While there can be a great deal of discussion during the training of scorers, the scorers need "handles" to make them more accurate and efficient -- key terms or phrases characterizing the different score points. If longer descriptions cannot be avoided, the highlighting of the key words or phrases is helpful.

Sometimes, one good approach to achieving brevity in score point descriptors is to focus primarily on the feature of a score point that distinguishes it from others without discussing much more -- i.e., highlight the "handles." The scoring rubric for Wilma's Wolves in Exhibit C does this. The question asks students to design an investigation. The 2-point response must manipulate the critical variable, the fullness of the moon. The 3-point response must control relevant variables. The 4-point response must address the issue of multiple trials or adequacy of sampling. The scoring rubric for Willard's Logic in Exhibit D indicates that the difference between the 2-point and 3-point responses is the generalizability of the argument. The "handles" illustrated in the rubrics for these questions include "CONTROL VARIABLES," "MULTIPLE TRIALS/SAMPLE SIZE," and "GENERALIZE." When the meanings of such terms are internalized by scorers, the scoring becomes fast and accurate. Scorers do have to look at other aspects of responses, but if they mentally determine initial scores for responses within seconds, and initial scores are usually the "correct" scores, then the remainder of scorers' time on responses can be brief. They simply have to confirm that there are no other conditions within the responses that would cause them to change their initial decisions.

UNITED WE STAND; DIVIDED WE FALL

The nature and quality of the scoring rubric for a question is as important as the question itself. The creator of one of these rubrics has the difficult task of finding a way to produce generalizable specificity. A rubric must "capture" the vast majority of responses students give to a question. It must be specific enough to capture the uniqueness of the question. This latter characteristic contributes significantly to scorer agreement rates and consistency of scoring over time (discussed earlier in conjunction with equating issues). One pitfall to be avoided is the use of a vague rubric and relying too heavily on training to convey the specific information that scorers need. This makes training more difficult and does not give the scorers the "handles" they need. For example, consider a reading passage describing specific movements and behaviors of a raccoon and an associated question that asks students how raccoons and humans are alike. A vague rubric for this question might say the following:

4 points: interprets the similarities between raccoon and humans
3 points: characterizes the similarities between raccoons and humans
2 points: identifies similarities between raccoons and humans
1 point: identifies a similarity between raccoons and humans

Only with extensive training do such descriptions of score points take on a common meaning across scorers. This vague rubric would create additional problems if it were to be released to the public. While a testing company is correct in explaining that without training, teachers cannot be expected to score students' responses exactly the way the testing company's trained scorers would, the rubric should at least be usable by the teacher as is, and should give the people viewing it confidence that the scoring of students' responses is not too subjective. In other words, a scoring rubric should stand alone. A more appropriate rubric for the question cited above might read as follows:

4 points: identifies general categories of similarities and supports them with specific examples
3 points: identifies general categories of similarities
OR
identifies a general category without examples and identifies specific examples of another type without identifying a general category
2 points: identifies a single general category or identifies several specific examples only
1 point: identifies one or two specific examples

General Categories	Specific Examples
intelligence	figures how to open jar
manual dexterity	uses utensils, opens jar
mimics human movements/behavior	sits at table

This is a rubric that could be used by a teacher without training or without his or her having to "start from scratch" in determining how responses to the question are to be scored.

Conclusions

A great range of scoring issues has been addressed in this paper -- from general issues about the scoring of different modes of performance-based assessment to specific issues about the characteristics of good scoring rubrics. The key points that have been made are:

1) On-demand open-response testing can be psychometrically sound, producing highly reliable results despite a lack of perfection in scoring;

2) Scoring consistency and test reliability are not the same;

3) Educators should distinguish among several modes of alternative assessments and not attribute difficulties associated with one to the others;

4) While training in the development of alternative assessments and scoring approaches may provide "formulas" for beginners, the goal for test developers should be to get away from rigid, formulaic approaches as soon as possible;

5) General scoring guides should be used with caution to guide the development of specific rubrics and to inform students of expectations;

6) It is human nature to look for universal rules for success in item and rubric development as well as test taking. Communications from testing companies and state education agencies should not encourage educators to believe such rules exist when they do not;

7) The scientific principle of parsimony applies to the development of scoring rubrics as well.

Scoring is an art. We can approximate good art by painting by the numbers, but should not lose sight of the distinction between authentic art and approximated art. Flexibility and creativity play an important role in preparing for the scoring of student work and ultimately in the effectiveness of the scoring process. Whether we are designing a whole scoring system or creating an item-specific scoring rubric, these are two attributes that will enable us to apply existing, sound measurement principles to the scoring of student performances.

References

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., Deibert, E. (1994). *Can Portfolios Assess Student Performance and Influence Instruction?* RAND, Santa Monica, CA.

Olson, L. (1995). The New Breed of Assessments Getting Scrutiny. *Education Week*, 14(26), 1, 10-11.

EXHIBIT A

GENERAL SCORING GUIDE

SCORE POINT 4	• You complete all important components of the question and communicate ideas clearly. • You demonstrate in-depth understanding of the relevant concepts and/or processes. • Where appropriate, you choose more efficient and/or sophisticated processes. • Where appropriate, you offer insightful interpretations or extensions (generalizations, applications, analogies).
SCORE POINT 3	• You complete most important components of the question and communicate clearly. • You demonstrate an understanding of major concepts even though you overlook or misunderstand some less-important ideas or details.
SCORE POINT 2	• You complete some important components of the question and communicate those clearly. • You demonstrate that there are gaps in your conceptual understanding.
SCORE POINT 1	• You show minimal understanding of the question. • You address only a small portion of the question.
SCORE POINT 0	• Your answer is totally incorrect or irrelevant.
BLANK	• You did not give any answer at all.

EXHIBIT B

MR. WALKER'S MATH EXAM

2. The math exam scores for the 21 students in Mr. Walker's homeroom were:

65 90 82 78 84 92 88 86 70 68 75
88 90 85 61 81 79 82 84 83 90

a. The mean or average of the above scores is 81. What is the median score? What is the mode?

b. Use the grid on page 4 of your answer sheet to make a bar graph showing the frequency or number of scores in each of the score ranges 60-64, 65-69, 70-74, etc.

c. Which is the best general indicator of this class's performance on the exam — the mean, median, or mode? Explain your answer.

NOTE FOR SCORERS — Rank-ordered scores are:

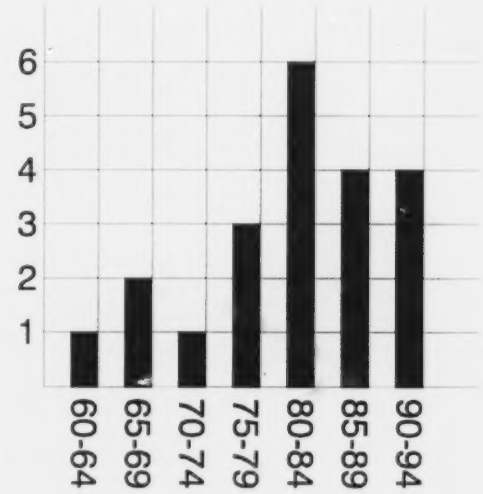
61 65 65 68 70 75 78 79 81 82 82 83 84 84 85 86 88 88 90 90 90 92

Total of 4 points can be awarded

1 point - correct value for median (83) OR correct value based on student's list with missing entry or two

1 point - correct value for mode (90) OR correct value based on student's list with missing entry or two

1 point - accurate graph as prescribed (minor error acceptable, e.g., miscount or one range omitted) (award ½ point for several minor errors or accurate line graph or bar for each score)



1 point - adequate defense of median as best general indicator in this instance (or if misdetermined median, good defense of mean; can award ½ for acknowledgment that median is most appropriate with weak explanation or if other parts of response indicate some understanding)

ENCODE SUM OF POINTS AWARDED; ROUND UP

0 points - none of the above or blank or irrelevant or other incorrect response

EXHIBIT C

WILMA'S WOLVES

Wilma has heard that wolves and their relatives (dogs) howl more at night when the moon is full. Describe in detail an investigation you could do to help you decide if this is true.

SCORING RUBRIC

4 points:	clearly describes data collection effort; varies fullness of moon; controls variables; involves MULTIPLE TRIALS (many nights) and/or more than one wolf/dog in design (SAMPLE SIZE)
3 points:	clearly describes data collection effort; varies fullness of moon; CONTROLS VARIABLES (e.g., makes sure weather and other conditions are the same); single trial or poor sampling
2 points:	clearly describes data collection effort; VARIES FULLNESS OF MOON (e.g., go outside when there is a full moon and listen and then do it again when there is not a full moon); does not control other variables
1 point:	describes data collection effort that would provide some relevant information; DOES NOT VARY FULLNESS OF MOON; may be slightly vague or unclear (e.g., outside when there is a full moon and listen)

EXHIBIT D

WILLARD'S LOGIC

QUESTION 1 — *Two students asked to have their math tests rescored. One student had a score of 50, which was far below the class average. The other had a score of 70, which was close to average. Willard believes that adding 20 points to the lower score would raise the class average more than adding 20 points to the higher score. Is Willard correct? Use the space on your answer sheet to explain or prove your answer to someone who disagrees with you.*

SCORING GUIDE:

3 Willard is NOT correct (stated or clearly implied) WITH

- correct algebraic proof OR
- correct explanation that the sum of the scores is the same either way OR
- specific counterexample shown or explained (must explain why it is a counterexample rather than merely state it)

NOTE: All of the above must generalize beyond the particular case.

2 Willard is NOT correct (stated or clearly implied) showing specific counterexample with some explanation that does not generalize

Other correct explanation that does not generalize

1 Willard is NOT correct (stated or clearly implied) with weak explanation only OR

Willard is NOT correct with specific counterexample with weak or no explanation OR

Willard is correct but with valid proof or explanation to the contrary

0 Willard is NOT correct with NO COUNTEREXAMPLE and incorrect or no proof/explanation OR

COUNTEREXAMPLE ONLY (i.e., no indication or implication of student's position relative to Willard's statement) OR

Willard is correct with incorrect or no proof/explanation OR

Blank or irrelevant or other incorrect response

NOTE: Stating the other position (that averages would be the same or that adding to the higher score would increase the average more) is NOT proof or explanation. It can be used to determine student's position relative to Willard's statement.